

DOCUMENT RESUME

ED 246 069

TM 840 340

AUTHOR Reid, Jerry B.
TITLE Adapting to a Mandated Pre-Set Pass/Fail Point.
PUB DATE Apr 84
NOTE 17p.; Paper presented at the Annual Meeting of the American Educational Research Association (68th, New Orleans, LA, April 23-27, 1984).
PUB TYPE Speeches/Conference Papers (150) -- Reports - Research/Technical (143)
EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Certification; *Cutting Scores; Policy; Scores; *Test Construction; Testing Problems; Tests
IDENTIFIERS Angoff Methods; Ebel Method; Nedelsky Method; *Standard Setting

ABSTRACT

While standard setting procedures are typically discussed in terms of deriving a reasonable cutting score for a given form of a test, the situation may be structured such that the standard has been mandated without regard to the test form itself. This situation may result either through legislative or policy actions and may be a fait accompli by the time someone experienced in standard setting methodology becomes involved. Although this may be an undesirable situation, it is not necessarily an impossible position from which to operate. This paper will explore an approach whereby an item "cut-score index" is included as an additional dimension in the test specifications for use in tailoring a test form to correspond to a pre-set pass/fail point. Issues such as the stability of such an item index and the effect on content representativeness are discussed. (Author)

* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *

ADAPTING TO A MANDATED PRE-SET PASS/FAIL POINT

Jerry B. Reid, Ph.D.
The American Registry of Radiologic Technologists
2600 Wayzata Boulevard
Minneapolis, MN 55405

"PERMISSION TO REPRODUCE THIS
MATERIAL HAS BEEN GRANTED BY

J. B. Reid

TO THE EDUCATIONAL RESOURCES
INFORMATION CENTER (ERIC)."

ABSTRACT

While standard setting procedures are typically discussed in terms of deriving a reasonable cutting score for a given form of a test, the situation may be structured such that the standard has been mandated without regard to the test form itself. This situation may result either through legislative or policy actions and may be a fait accompli by the time someone experienced in standard setting methodology becomes involved. Although this may be an undesirable situation, it is not necessarily an impossible position from which to operate. This paper will explore an approach whereby an item "cut-score index" is included as an additional dimension in the test specifications for use in tailoring a test form to correspond to a pre-set pass/fail point. Issues such as the stability of such an item index and the effect on content representativeness are discussed.

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION
CENTER (ERIC)

- ☒ This document has been reproduced as received from the person or organization originating it.
- ☐ Minor changes have been made to improve reproduction quality.
- Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

Paper presented at the annual meeting of the American Educational Research Association, New Orleans, April 1984.

Introduction

While it may seem obvious to those working in testing that a pass/fail point set in the absence of familiarity either with the specific items on the test or at least a set of well defined objectives from which the test is developed is not a reasonable approach, experience shows that this is not always obvious to others. The simple example that a score of 70% on a test form made up of relatively easy items implies a different level of knowledge than a score of 70% on a more difficult set of items covering the same content often clarifies the underlying problem with establishing a pass/fail point not referenced to specific examination content to someone unfamiliar with standard setting. However, this may not adequately address the sociopolitical issues responsible for the movement to establish a given standard in the first place.

While the "perception of propriety" of a pass/fail point is not of technical concern, it is probably the second most important concern (next to content coverage) influencing the acceptance of a test by the target community. This perception of propriety is of particular importance in certification and licensing programs since not only must the end users of the test results accept the standard, but also the examinees (who are usually adults for certification/licensing situations) must perceive the standard as being appropriate.

Although the general issue of mandated pass/fail points applies to any test on which a standard is set, this discussion will be limited to certification and licensing tests.

Origin of the Problem

There are at least two ways in which measurement specialists may find themselves in the situation where a pass/fail point has been mandated without reference to the specific items on the test form. The first situation arises through influence by the profession involved and/or the consumers of the service provided by the profession. Certification organizations are typically structured in one of two ways. Under the first structure, the certification organization is controlled directly by the professional group, being a subdivision of the organization. Under the second structure, the certification organization is administratively independent, but policy is set by a board composed in whole or in part of representatives appointed or elected by the professional organization. Either structure may provide an avenue whereby influence is exerted to establish a given pass/fail point on the certification examination. There is a good chance of avoiding the mandated standard problem provided that someone knowledgeable in the issues of standard setting becomes involved before policy decisions are made. There are many certification organizations, however, that either work on a limited basis with testing service organizations or that rely on consultants only infrequently.

The second way in which a mandated standard may come about is through legislative channels. States requiring that certain occupations be restricted to licensees, often make use of the test results from national certification agencies rather than establishing their own testing services. While such an arrangement may involve substantial interaction between the state and the certification agencies, it is not uncommon for the arrangement to be non-interactive. In such cases the

state may have no direct contact with the certification agency, but simply require those persons wishing to practice in the state to send proof of their certification by the national organization. States developing licensing legislation for an occupation based upon such a non-interactive model may develop legislation (or revise existing legislation) without seeking input from the certification agency. Examples exist where laws were developed in this way and a capriciously established pass/fail point written directly into the law. Such legislative activity may not come to the attention of the certification agency until too late. On the other hand it may come to the certification agency's attention, but the simplicity provided by an arbitrary standard may make the agency's recommendation for not establishing such standards politically unpalatable. It is much easier to rally around the battle cry of "70% correct or no license" than that of "a standard set by an appropriate method or no license."

Benefits Versus Risks

While discussing the issue of standard setting as researchers, it may be desirable to think in terms of setting a pass/fail point for a specific test form. As a practical matter, however, it is worth considering the feasibility of starting with a pass/fail point and tailoring a test to fit the standard. Prior to discussing the mechanics involved, an analysis of the potential benefits and risks is in order.

The primary benefit of the approach is the possibility of providing a test with a non-capriciously established "standard" for a situation in which the actual cut score has been capriciously set. Used in the sense above, "standard" refers to the underlying scale of competency as

distinguished from the scale provided by the items used on a particular form of the test.

Whereas in some situations adapting to an capriciously set pass/fail point may be a necessity, in other cases it may be proposed to enhance the "perception of propriety." The situation is analogous to that of the face validity of tests. While face validity is not accepted as evidence of a test's appropriateness for a given use, it is conceded that unless a test appears to cover appropriate material, the acceptance of the test by the community will be uncertain. Similarly, while the perception that a given standard is appropriate is not evidence that it is reasonable, a standard that lacks the appearance of propriety may negatively affect the reputation of an otherwise well-respected instrument.

The procedure runs the risk of creating the impression that the standard is being somehow "fudged" and that the original intent of the mandated standard is being undermined. Although this is a very real concern, in those situations where such a standard exists and cannot be modified, running this risk may be the only practical alternative.

Perhaps of more concern is the possibility of misuse and/or misinterpretation of a standard set by this procedure. A test tailored to a pass/fail point of 80%, for example, could be compared to an examination covering the same content but having a pass/fail point of 70% and mistakenly touted as having a more stringent standard while in fact the difference simply reflects the difference in difficulty levels of the items composing the tests.

In light of the potential risks, it would seem appropriate that consideration of the method of tailoring a test to a given pass/fail point be restricted only to those situations where the pass/fail point

has been mandated and cannot be changed. In this situation the only option may be to use a test with a truly capricious standard, to not provide any test, or to attempt to adapt to the mandated pass/fail point.

Serious potential technical pitfalls are associated with the approach of "backing into" a mandated standard as well. The overriding concern is naturally that of adversely affecting the reliability and/or validity of the test. As an extreme example, if the "acceptable" standard is 100%, it would not be feasible to tailor a test to this standard. At the other extreme, a standard set very low would require, for the typical certification situation where some part of the examinees are in truth failers, the use of very difficult items in order that part of the distribution would fall below the pass/fail point. Selection for such extreme difficulty would be very likely to result in an overabundance of items that tended to be less relevant. Although all items in the item bank are assumed to be relevant, the items would range from the very relevant to the marginally relevant. For the situation in which training programs are well established, graduates have been presented the appropriate training, and there is a good match between the ability of the students selected into the program and the occupation's requirements, the most relevant items should be relatively easy for the majority of examinees. Items that are difficult for this group are often, although not always, less relevant. For the typical certification situation a very low pass/fail point would require the use of the least relevant items from the bank.

Assuming that the "acceptable" pass/fail point is not extreme and that the item bank has an adequate supply of relevant items with a range of item difficulties, how might it be possible to tailor a test to a mandated pass/fail point?

Proposed Method

The method being proposed for "backing into" a mandated pass/fail point is straightforward, but does require that the test development situation be such that a bank of items exists from which a form of the test is to be developed (as opposed to the situation in which the test form already exists at the time the pre-set standard is mandated). The method requires that a "cut-score index"¹ be assigned to each item in the bank. This index would be used as an additional dimension in the selection of items for inclusion on a test. Not only would it be necessary to select a given number of items from each content category, but it would also be necessary that the mean cut-score index for the items selected be equal to a given value.

The method logically restricts the choice of a standard setting method to those in which judgements to individual items are made. The most commonly used methods in this category are the Nedelsky (1954), Angoff (1971), and Ebel (1972) Methods. Before discussing the use of these methods, a possible short-cut method will be discussed.

Method #1 (A Possible Short-Cut)

To the extent that the item judgements under the standard setting methods are a reflection of item difficulty, it would be possible to utilize the item difficulty index for roughly tailoring a test to a given standard. Various researchers have reported data on the relationship of item difficulty and the cut-score index derived from the

¹ The index is referred to as an "item cut-score index" in this paper rather than previously used terms that are associated with a specific standard setting method (e.g., MPL for the Nedelsky Method) to emphasize the generic character of the index.

standard setting methods. While not strong, a relationship has generally been shown. (Gross, 1983; Guerin & Smilansky, 1976; Taylor, Reid, Senhauser & Shively, 1971). Van der Linden (1982) discussed this relationship in terms of inconsistent probabilities (i.e., low probabilities for easy items, high probabilities for hard items). Van der Linden's conceptualization of these as "inconsistencies" is based upon the premise that the judgements are being made solely upon the item's difficulty. While the Ebel Method specifically separates judgements on items into difficulty and relevance factors, neither the Nedelsky nor the Angoff Methods do. The typical instructions to judges for these methods do not make specific mention of how or whether an item's relevance should affect the judgement. As discussed by Livingston and Zieky (1982), "the judges' task is to look at the question and identify the wrong answers that a borderline test taker would be able to recognize as wrong" for the Nedelsky Method. For the Angoff Method the judges' task is to "make a judgement of the probability that a borderline test taker would answer the question correctly." The key word in both of these descriptions is "would." This implies a judgement regarding actual performance on the item based upon an item's difficulty for the borderline examinee devoid of any considerations of the item's relevance. The calculations for deriving the pass/fail point treat the judgements as probabilities of correct response and are therefore consistent with the directions given.

In working with one group of judges, it was brought to my attention while discussing the ratings that had been made that although some judges had made their ratings based primarily upon the difficulty factor, others took into consideration the item's relevance as well. Those taking into account the item's relevance might assign different

cut-score indices to two items having the same perceived difficulty. The item that was extremely relevant would be assigned a higher cut-score index than another item having the same perceived difficulty, but being less relevant. Although in reality, the borderline group might have the same probability of correctly answering both items, the one item was judged to be more important to answer correctly to be classified as a passer. While all items on a certification or licensing test should be relevant, there will always be varying degrees of relevance. Discrepancies between the difficulty index and the cut-score index do not necessarily reflect inconsistencies in judgements, but rather may result from multidimensional decisions.

To the extent that relevance influences the judgements, the item's difficulty should not be expected to be perfectly related to the cut-score index. The difficulty index will provide only a very imperfect way of tailoring a test to a given standard. A clarification of instructions to judges on this point may help to reduce variability of judgements on individual items and increase the observed relationship between item difficulty and cut-score indices.

Method #2

The most direct way of backing into a mandated standard involves the use of cut-score indices assigned to all items in the bank with either the Nedelsky, Angoff or Ebel Method. The use of cut-score indices in this way assumes a level of precision and stability that is over and above that required of the standard setting methods when used under normal conditions. When the standard setting methods are used as originally proposed, it can be assumed that various sources of error

balance out when the indices are averaged across items to arrive at the overall cut score. Since only the total test cut score is of interest, the judgement of individual items and the associated errors have tended to be ignored. The focus on the judgements assigned to individual items has the potential to be more sensitive to these errors. As an example of one such potential source of error is the effect of the characteristics of items surrounding the item being judged. It is possible that an item surrounded by very easy, very relevant items may appear to be more difficult and less relevant than the same item when it is surrounded by very hard, less relevant items. While within the same test these effects may balance out (since the item may have roughly an equal and opposite effect on the items around it), the index assigned to the item would be a characteristic not just of the item itself, but also the environment created by the surrounding items. When the item appeared next to other items, its previously determined index would no longer hold.

Each of the three methods will be briefly discussed as they relate to the proposed method. Familiarity with the methods will be assumed in the discussion.

Nedelsky Method: As discussed by Brennan and Lockwood (1980), van der Linden (1982), and Shepard (1980), the ratings possible for a item under the Nedelsky Method are restricted by the number of options. For a four option item, the possible ratings are 0.25, 0.33, 0.50, 1.0. The gap between 0.50 and 1.0 is especially disturbing. If ratings are made by arriving at a group consensus, then the consequences for a individual item's cut-score index is greatly affected by this coarseness of scale. If cut-score indices are derived by averaging the individual ratings of

judges, the effect of this coarseness is mediated somewhat, but not eliminated. To the extent that the borderline examinee is not expected to answer all items correctly, this gap tends to force the individual cut-score indices towards 0.50. This in turn tends to depress the overall standard on the test. This potential bias towards lower cut-score indices with the Nedelsky Method is a limitation of the method for this as well as normal applications.

Angoff Method: Since the Angoff Method assigns a value to the overall item on a continuous or at least equally spaced scale rather than to individual options, it avoids the problem associated with the Nedelsky Method. This advantage, however, also leads to a possible disadvantage. Since a close analysis of the distractors is not a requirement for making judgements (although this is an intended activity) judges may have a tendency to ignore the effect of the distractors on item difficulty in making their judgements. If the particular distractors used make the item more difficult than the stem alone suggests, this may lead judges to underestimate item difficulty. This may in fact partly explain why the Angoff Method tends to lead to the setting of higher standards than does the Nedelsky Method.

This potential bias towards higher cut-score indices by the Angoff Method is a possible limitation of the method for use in determining cut-score indices for individual items.

Ebel Method: There are two types of judgements required by the Ebel Method. The first involves judgements regarding the percentage of items from each classification that will be answered correctly by the borderline examinee. The second judgement involves assigning items to

the various classifications based upon the item's difficulty and relevance. A pre-set standard could be "backed-into" under this method by taking the percentages assigned to the classifications as a given, and selecting items from the various classes so as to total to the pre-set standard. A reasonable precaution would be to select items from across the scheme, relying on the most relevant items as much as possible to arrive at the desired standard. A potential short-coming of this method is the requirement that judgements be made on two dimensions, especially when there is some doubt regarding judges' capability to make judgements based on a single dimension.

Composite Method: While it is still unknown as to which factors affect the judgements under each of the methods, it seems prudent to base item cut-score indices upon composites (averages) of the results from two or more methods (as is typically recommended when the methods are used in normal situations). Individual judgements resulting from the Nedelsky and Angoff Methods are the easiest to combine and represent a logical choice given the usual finding that Angoff results in a higher standard and Nedelsky in a lower standard.

General Considerations

Representativeness of Judges: While much has been said regarding the finding that judges using the same method sometimes arrive at different standards, this does not seem surprising given that the conceptualization of borderline performance will vary legitimately across judges. This is particularly true of certification/licensing examinations on which the perception of borderline would be expected to

differ across representatives from the various interested parties (educators, entry-level job incumbents, supervisors, administrators, consumers, et cetera). The important point is that input from all of these groups should be considered for each item in the bank (provided that the judges are qualified to make the judgements asked of them).

Number of Judges: While the number of judges has tended to be small for practical reasons (usually the number readily available), the numbers may well be too small given the importance of the decisions being made. While an item difficulty index based upon ten examinees would not be expected to be very meaningful, cut-score indices are typically set with half that number of judges. The ratings made by the judges are probably a more complex intellectual activity and therefore subject to more sources of error than is selecting the correct answer to an item. Although any method of standard setting would benefit from having many judges, the assignment of a cut-score index to all items in the bank would make this especially important given the use proposed for the indices.

Across-Form Consistency: While setting a standard for a base form of the test, and statistically equating future forms back to the base form to maintain a constant standard is preferred, some groups either due to the small number of examinees or other constraints, set a new standard on each form using one of the standard setting methodologies. The approach of assigning a cut-score index to each item in the item bank and assembling forms using these indices could save such groups considerable time in the long run.

Verifying Standards on Assembled Form: The question arises as to whether a test form assembled based upon the individual cut-score indices should undergo a standard setting procedure as a whole. To the degree that the original individual item indices were contaminated by the surrounding items, a new standard, at variance from the desired standard, might result from such an exercise. The problem of resolving differences must then be addressed. The number and representativeness of judges used to establish the original item cut-score indices as opposed to the judges reviewing the assembled test form should be taken into consideration in deciding how to resolve differences.

If a separate panel is to rate the items on the assembled test form, extra items should be included with a range of cut-score indices so that items may be selectively replaced to arrive at the desired standard.

Conclusion

For those who subscribe to the position put forth by Glass (1978) that trying to establish absolute standards is futile given the inherent difficulties, the approach just described will be viewed as an attempt to raise futility to new heights (or depths).

Those who are convinced of the necessity for standard setting may feel that the approach is asking too much of the admittedly imperfect methods that are available for setting standards. The approach may, however, help standard setters to examine more closely some of the factors that affect the ratings of items and perhaps begin to the resolve why results from the various methods differ.

While adaptation to a mandated standard represents an additional complication to the already complex situation of determining a

defensible standard, situations may arise in which no practical alternatives exist. In addition to the situation in which a standard has been mandated, tailoring a test to a given standard may have benefits in terms of satisfying the test users' "perception of propriety." Great care should be exercised if used with the latter intent since the potential exists for misinterpretation and mis-use of a standard derived in this manner.

REFERENCES

- Angoff, W. "Scales, norms and equivalent scores." In R. L. Thorndike (Ed.) Educational Measurement (2nd ed.). Washington: American Council on Education, 1971.
- Brennan, R. L. & Lockwood, R. E. "A comparison of the Nedelsky and Angoff cutting score procedures using generalizability theory." Applied Psychological Measurement, 1980, 4, 219-240.
- Ebel, R. L. Essentials of Education Measurement. Englewood Cliffs, NJ: Prentice-Hall, 1972.
- Glass, G.V. "Standards and criteria." Journal of Education Measurement, 1979, 15, 237-261.
- Gross, L. J. "A refinement in the Nedelsky procedure for setting cutoff scores on credentialling examinations." Paper presented at the National Council on Measurement in Education, Montreal, April 1983.
- Livingston, S. A. & Zieky, M. J. Passing Scores: A Manual for Setting Standards of Performance on Educational and Occupational Tests. Princeton, NJ: Educational Testing Service, 1982.
- Nedelsky, L. "Absolute grading standards for objective tests." Educational and Psychological Measurement, 1954, 14, 3-19.
- Shepard, L. "Standard setting issues and methods." Applied Psychological Measurement, 1980, 4, 447-467.
- Van der Linden, W. J. "A latent trait method for determining intrajudge inconsistency in the Angoff and Nedelsky techniques of standard setting." Journal of Educational Measurement, 1982, 19, 295-308.